Understanding face representations in deep convolutional neural networks: Face space theory evolves

Alice J. O'Toole

The University of Texas at Dallas

12/12/18

Acknowledgement

- UTD lab
  - Asal Barachizadeh
  - Jackie Cavazos
  - Ivette Colon
  - Matthew Q. Hill
  - Carina Amanda Hahn
  - Ying Hu
  - Gerie Jeckeln
  - Eilidh Noyes
  - Kim Orsten-Hooge
  - Connor Parde
  - Diana Zi
  - Univ. of Maryland (UMIACS)
    - Carlos Castillo
    - Rama Chellappa
    - Jun-Cheng Chen
    - Swami Sankaranayaran
    - Rajeev Ranjan
  - Univ. of Siegen
    - Volker Blanz
- NIST
  - P. Jonathon Phillips

Lab Research funded TSWG/DOD, NIJ, NIST, IARPA JANUS Program

Parde, Hill, Colon, Castillo, Chen, Sankaranayraran & O'Toole, A. J. (2017). *Proceedings of the IEEE Conference on Automatic Face and Gesture Recognition*

O'Toole, Castillo, Parde, Hill & Chellappa (2018). Face space representations in deep convolutional neural networks. *Trends in Cognitive Science*.

Hill, Parde, Castillo, Colon, Ranjan, Chen, Blanz & O'Toole, (*submitted*). Deep convolutional neural networks in the face of caricature: Identity and image revealed.

Acknowledgements. Work supported in part by the Intelligence Advanced Research Projects Activity (IARPA). This research is based upon work supported by the Office of the Director of National Intelligence (ODNI), Intelligence Advanced Research Projects Activity (IARPA), via IARPA R&D Contract No. 2014-14071600012. The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements, either expressed or implied, of the ODNI, IARPA, or the U.S. Government. The U.S. Government is authorized to reproduce and distribute reprints for Governmental purposes notwithstanding any copyright annotation thereon.

Overview

- Face recognition over changes in image & appearance
  - human ability

- face representations in deep convolutional neural networks (DCNNs)

How many identities here?

Jenkins et al. (2011)





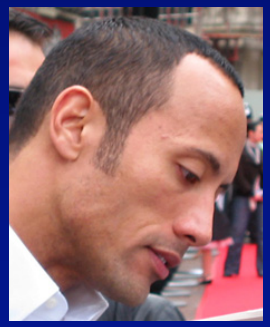



Hallmark of Face Familiarity

- generalize identification over change in:
  - pose
  - illumination
  - expression
  - appearance
- ***Progress*** - face recognition software > 2014
  - Deep convolutional neural networks (DCNNs)

“In-the-wild” images

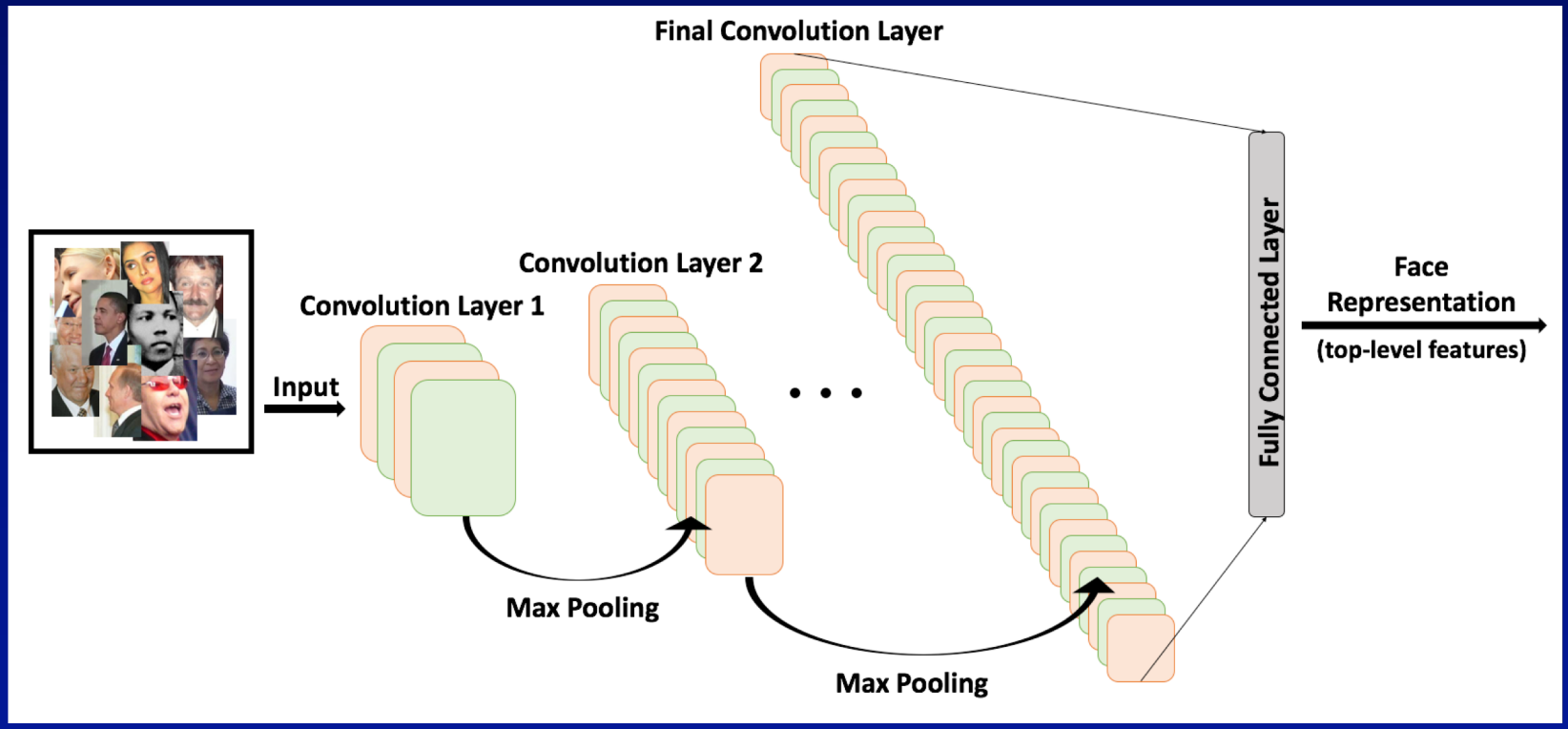
- DCNNs effective over large variations in viewpoint and illumination and expression (PIE)




Deep Convolutional Neural Nets (DCNNs) (Krizhevsky, et al., 2012)

- multiple layers of simulated neurons
  - convolve and pool image data
    - representations expand in early and intermediate layers
  - small number of fully connected layers
    - representations compressed in top layer
  - final representation of image emerges at top layer
    - *highly compact*

Deep Convolutional Neural Nets (DCNN)



Face Identification (DCNNs)

- state-of-the-art in automatic face identification
  (Ranjan et al., 2018; Taigman et al., 2014; Parkhi et al., 2015; Schroff et al., 2015)
- "top-level" face representation

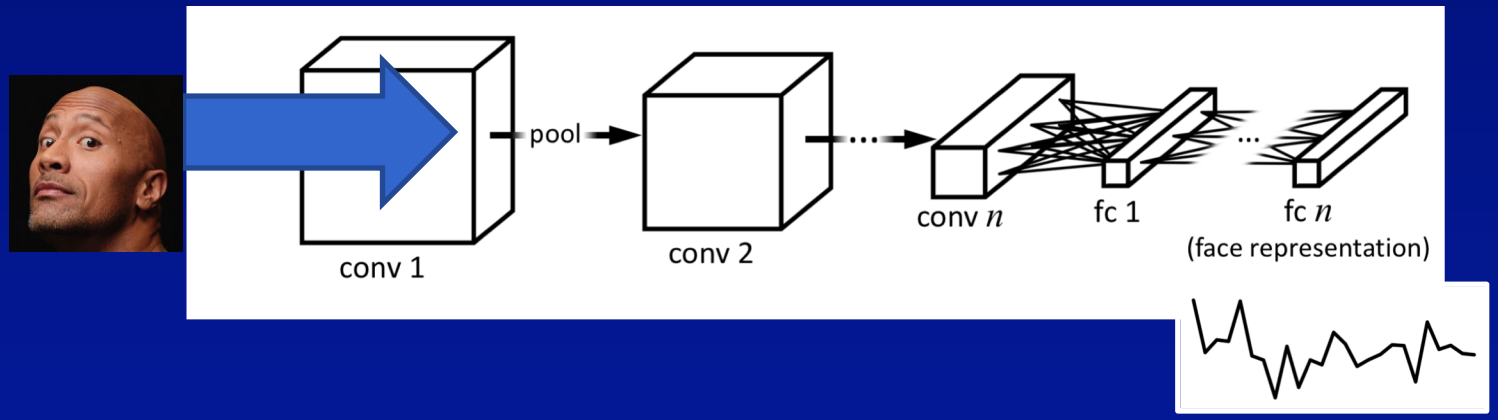


image credit
"Dwayne Johnson" by fmovies st (public domain)

Representations used for face identification

Understanding the DCNN Code

Why this is challenging:

- number of nonlinear computations
- uncontrolled nature of training data

Number of non-linear operations

- *Why* do they work?
  - 39.8 million computations between image and representation!

NETWORK B

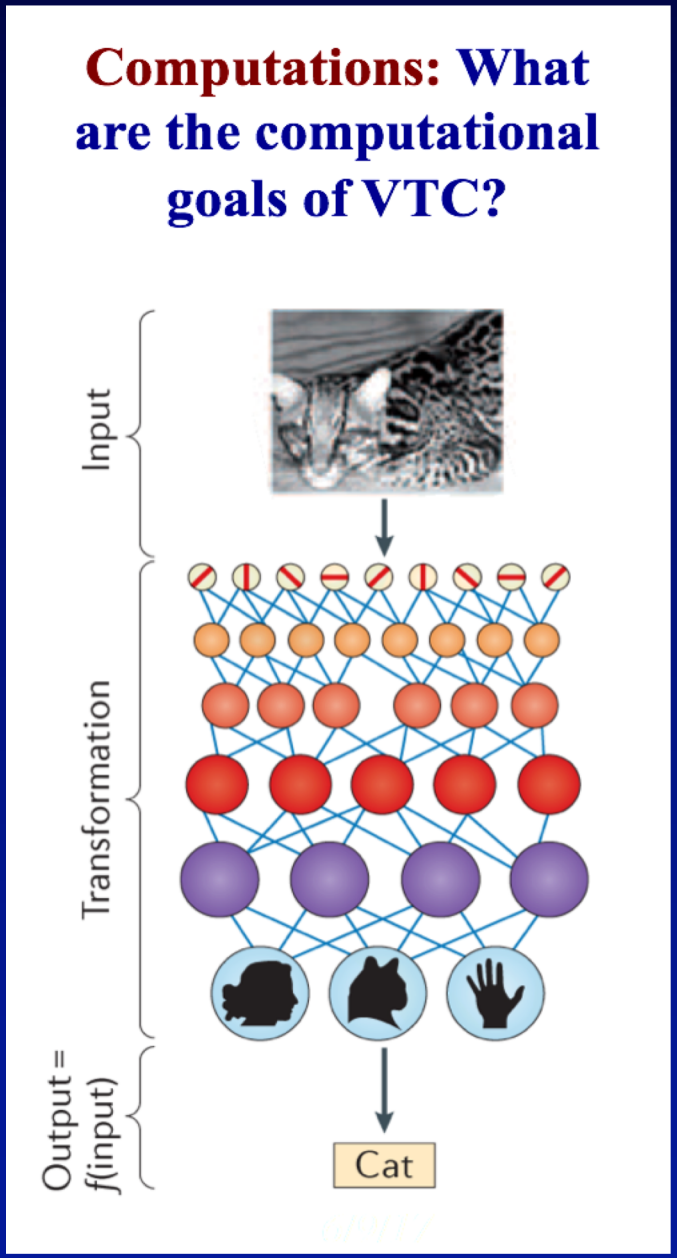
Layer	Kernel Size/Stride	Parameters
conv1	11 x 11/4	35K
pool1	3 x 3/2	
conv2	5 x 5/2	614K
pool1	3 x 3/2	
conv3	3 x 3/2	885K
conv4	3 x 3/2	1.3M
conv5	3 x 3/1	2.3M
conv6	3 x 3/1	2.3M
conv7	3 x 3/1	2.3M
pool7	6 x 6/2	
fc6	1024	18.8M
fc7	512	524K
fc8	10548	10.8M
Softmax Loss		Total 39.8M

Sankaranayaran et al. (2015)

Face representation 512 element vector

Visual Cortex

- Why does *it* work?
  - DCNNs designed to model primate visual system
  - 100's of millions computations between image and categorical representation in inferotemporal cortex!



Grill-Spector et al., 2014

Uncontrolled Image Training

Network Training
Web-scraped in-the-wild images

- millions of face images
- 10's of thousands of identities

- No control over:
  - viewpoint, illumination, occlusion, image quality, etc.
  - number of images per identity
  - diversity of images per identity

Goal

- understand face representations in DCNNs that achieve robust face recognition

- *Why is this important?*
  - face recognition software makes errors
  - anticipating these errors requires understanding

Approach

- visualize the similarity "space" of face representations
  - highly structured

- probe an "uncontrolled" network with controlled images and visualize representations of image variables

Approach

- visualize the space of face representations
  - highly structured
    - (Parde et al., 2017)

- probe the network with controlled images and visualize representations of image variables
  - Hill et al. (in prep.)

Networks

- analyze top-level features from state-of-the-art DCNNs:
  - Network A (Chen et al., 2015)
  - Network B (Sankaranarayanan et al., 2016)
- developed for IARPA Janus Competition
- trained on CASIA Webface database
  - 490,000+ images, 10,000+ identities
- top-level feature descriptor length:
  - Network A –> 320 features
  - Network B –> 512 features
- Test set: 25,787 images of 500 identities

NETWORK A

Name	Filter Size/Stride	Output	Parameters
conv11	3x3x1/1	100x100x32	.28K
conv12	3x3x32/1	100x100x64	18K
pool1	2x2/2	50x50x64	
conv21	3x3x64/1	50x50x64	36K
conv22	3x3x64/1	50x50x128	72K
pool2	2x2/2	25x25x128	
conv31	3x3x128/1	25x25x96	108K
conv32	3x3x96/1	25x25x192	162K
pool3	2x2/2	13x13x192	
conv41	3x3x192/1	13x13x128	216K
conv42	3x3x128/1	13x13x256	288K
pool4	2x2/2	7x7x256	
conv51	3x3x256/1	7x7x160	360K
conv52	3x3x160/1	7x7x320	450K
pool5	7x7/1	1x1x320	
dropout (40%)		1x1x320	
fc6		10548	3296K
softmax cost		10548	
total			5006K

(Chen et al., 2015)

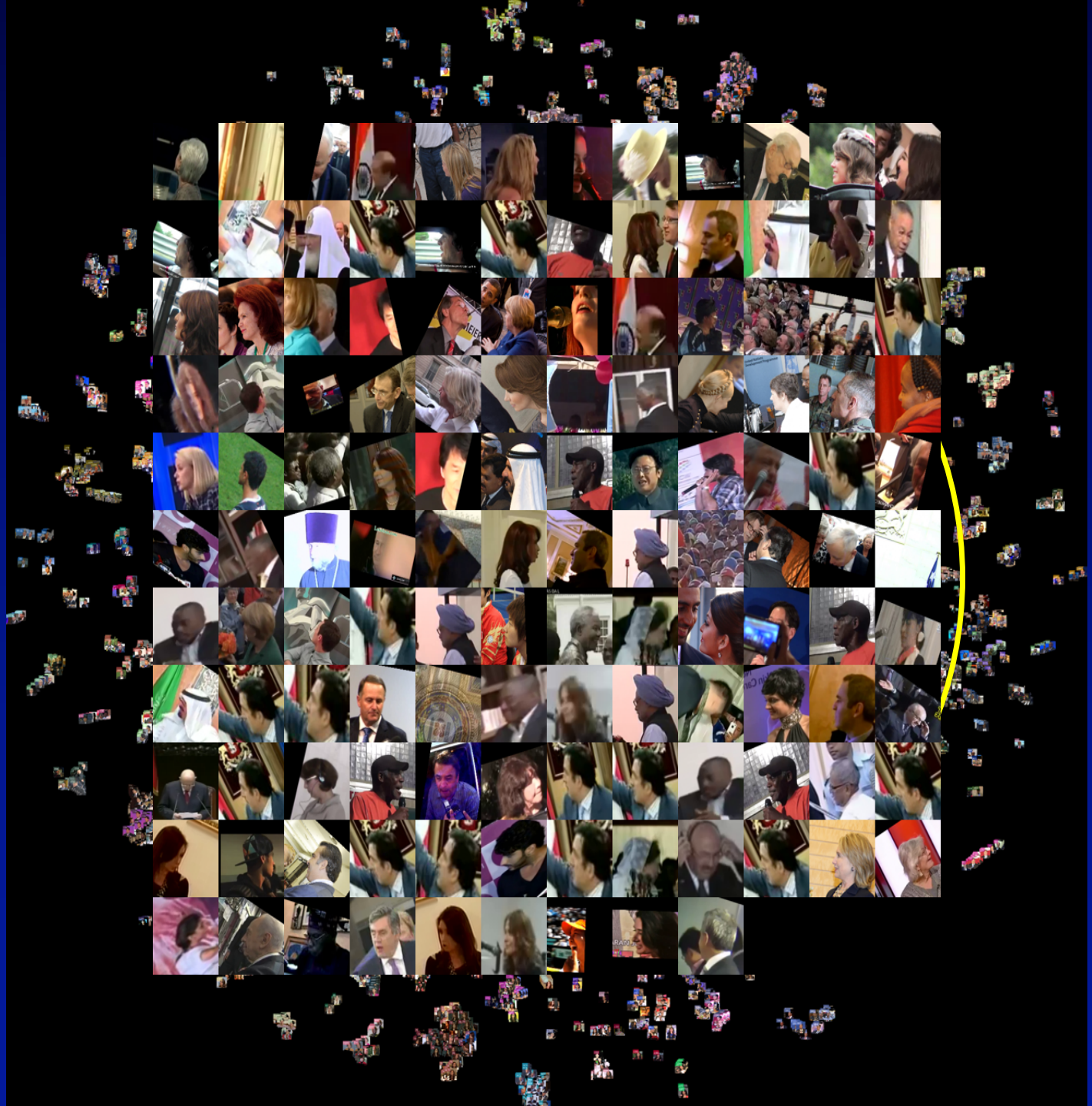
NETWORK B

Layer	Kernel Size/Stride	Parameters
conv1	11 x 11/4	35K
pool1	3 x 3/2	
conv2	5 x 5/2	614K
pool1	3 x 3/2	
conv3	3 x 3/2	885K
conv4	3 x 3/2	1.3M
conv5	3 x 3/1	2.3M
conv6	3 x 3/1	2.3M
conv7	3 x 3/1	2.3M
pool7	6 x 6/2	
fc6	1024	18.8M
fc7	512	524K
fc8	10548	10.8M
Softmax Loss		Total 39.8M

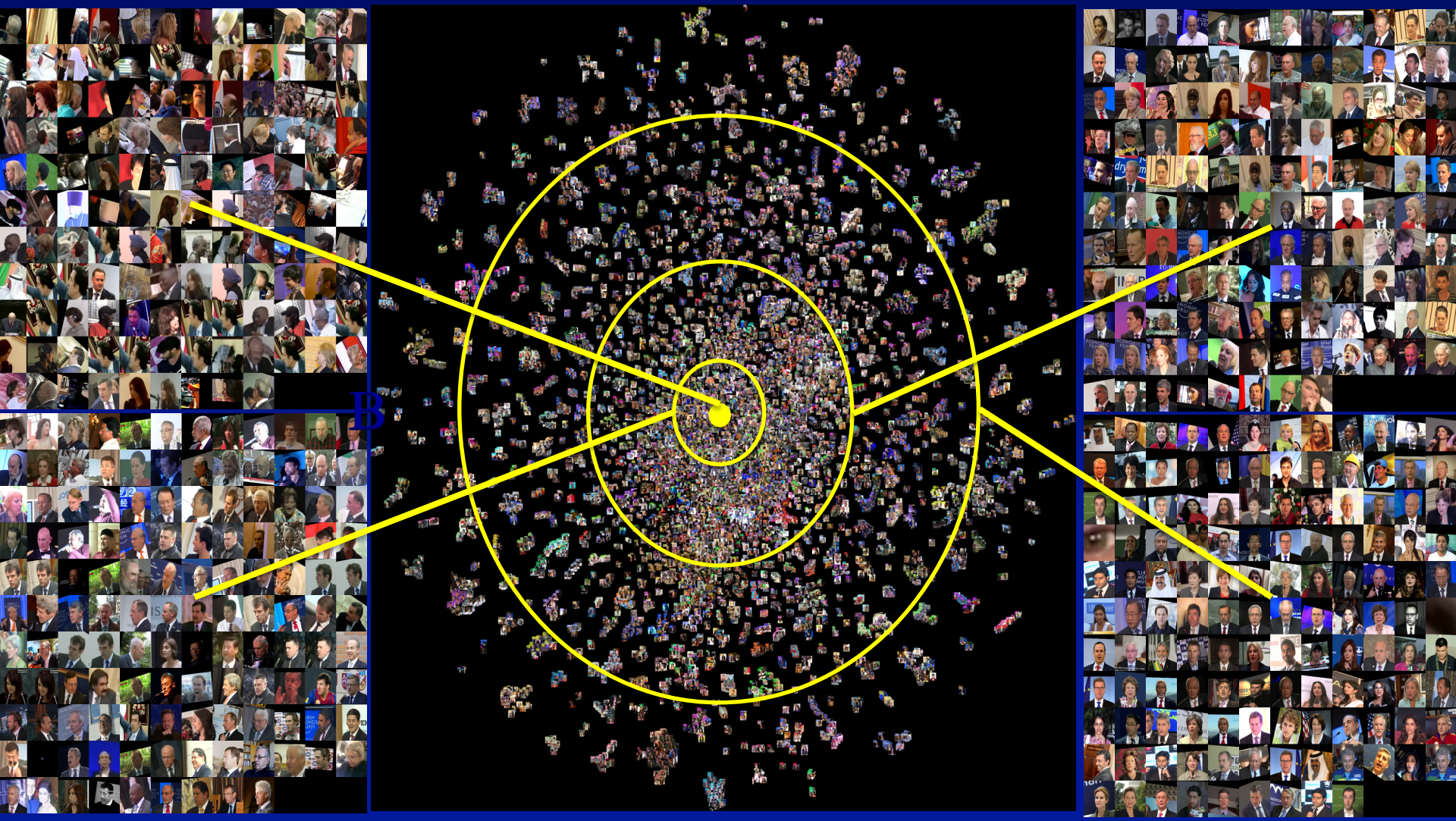
(Sankaranarayanan et al., 2016)

Visualization

- Face similarity space
  - structure of the space of 25K in-the-wild images

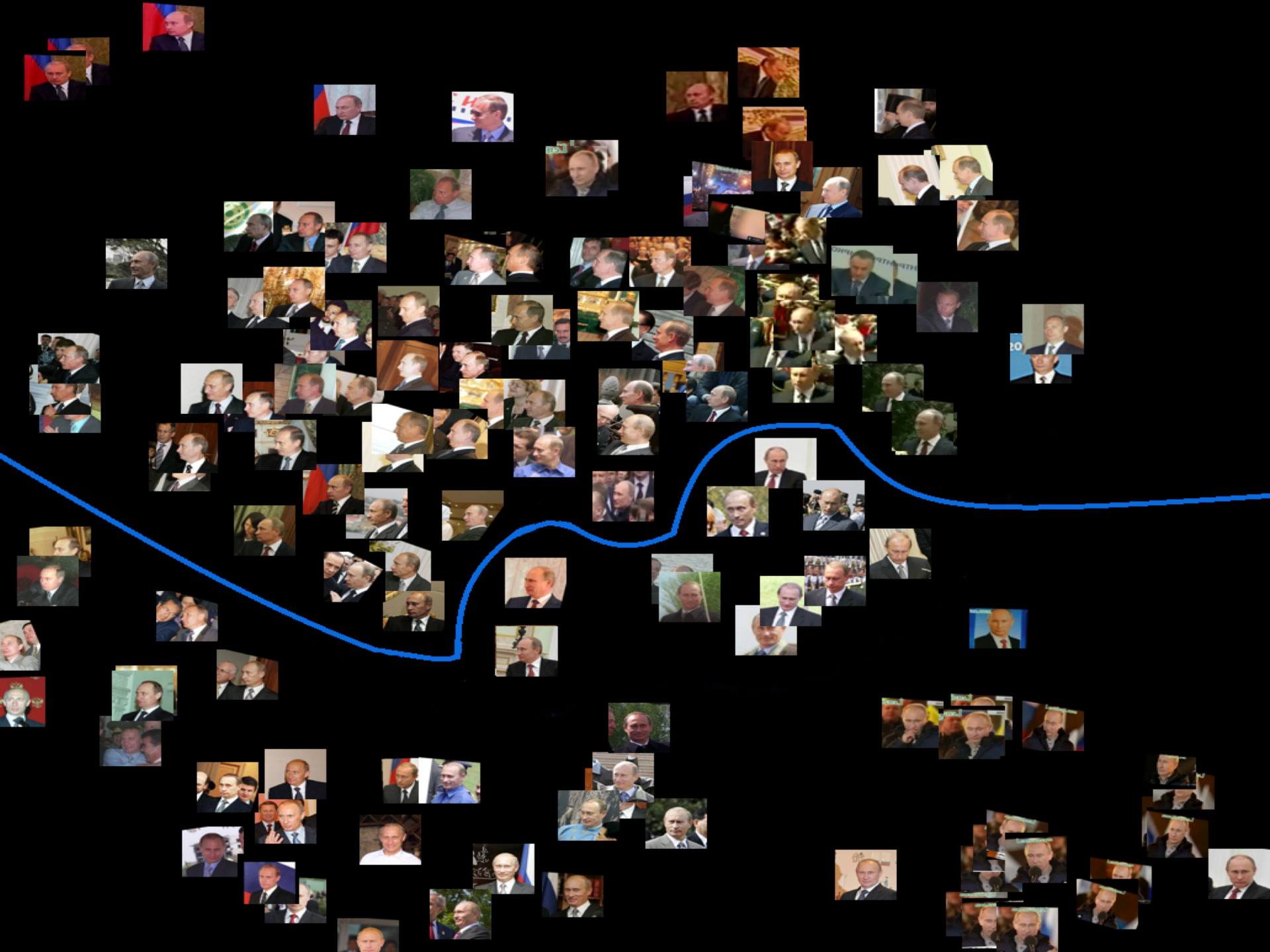


Structure of the DCNN Space Codes Image Quality



And also, viewpoint

- visualized 1 identity in a face space made from the top-level features of DCNN
- 1 identity = many images and videos
- embedded in visualization of 25K test images



Assumption

- to achieve robust face recognition DCNNs
  - create a representation that transcends the image data
  - deletes or compensates for nuisance variables
    - view
    - illumination
    - etc.
  - representation eliminates image data to extract identity

Not what DCNNs do.

Goal

- investigate **organization** of information encoded in top-level representation
  - identity
  - gender
  - viewpoint
  - illumination

- Approach
  - -> probe “in-the-wild” network with “in-the-lab” dataset

Two types of image data

- "in-the-wild" images
  - often web-scraped
  - unbalanced (identity, viewpoint, etc.)
  - unconstrained viewing conditions

- "in the lab" controlled laser-scanned faces (Blanz & Vetter, 1999)
  - can be rendered under precise, arbitrary conditions

Experiments

- 1 - create DCNN "face similarity space" from uncontrolled images (Chen et al., 2015; Ranjan et al., 2017)
  - project controlled images into the face space
  - "see the DCNN representation of image and identity features"

- 2 - apply morphing to manipulate facial distinctiveness
  - project **caricatured** images into the uncontrolled space
  - understand representation of identity in DCNNs

Face identification DCNN

(Chen, Patel, & Chellappa, 2016)

- identity training (Network A)
  - 494,414 face images of 10,575 identities
  - CASIA-WebFace dataset (Yi, Lei, Liao, & Li, 2014)

- training images varied widely in
  - illumination
  - viewpoint
  - quality (blur, facial occlusion, etc.)

Face Identification Performance

(Chen, Patel, & Chellappa, 2016)

- state-of-the-art in 2016
  - developed for IARPA Janus competition
  - performs well on IJB-A dataset (Klare et al., 2015)
  - performs as well as forensically trained fingerprint examiners on a challenging test of face identification (Phillips , et al., 2018)

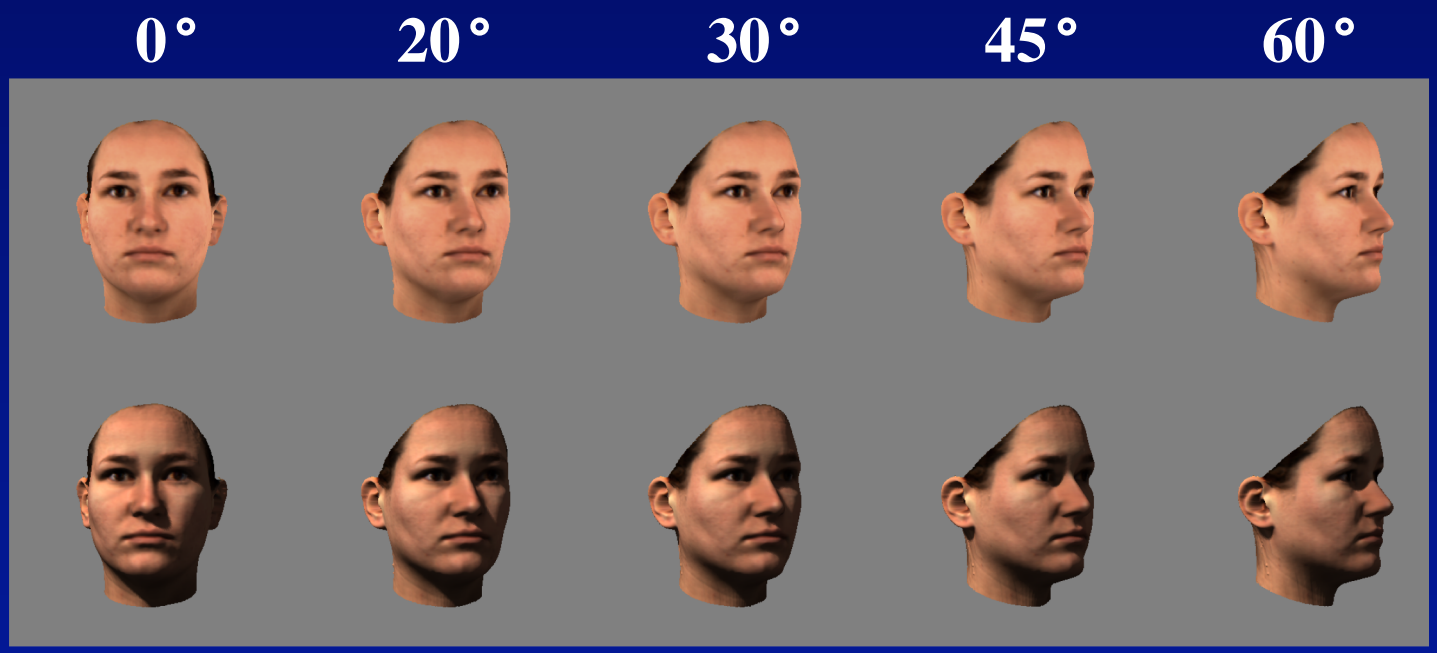
Controlled Face Images

- MPI Faces dataset (Troje & Bülthoff, 1996)

- 133 laser-scanned face identities
  - 65 male, 68 female

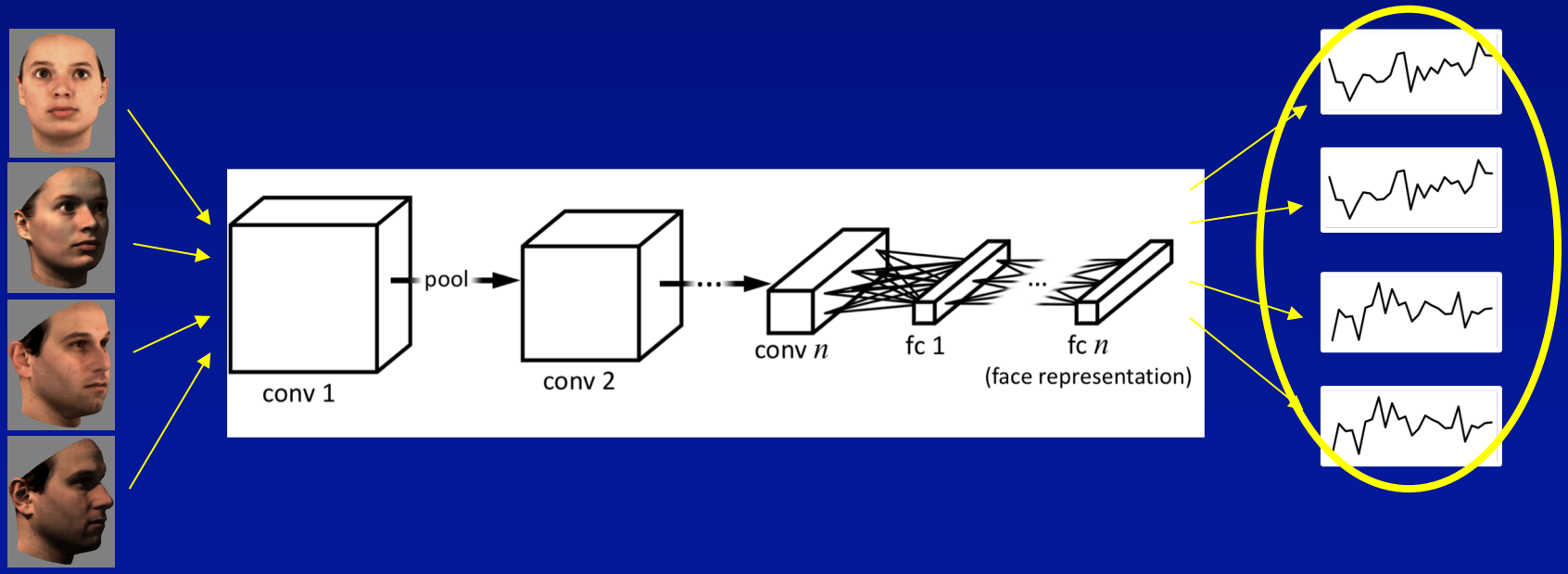
- laser-scan allows rendering varied systematically in:
  - viewpoint (5 levels from frontal to profile)
  - illumination (2 levels: ambient, spotlight)

Example Face Identity



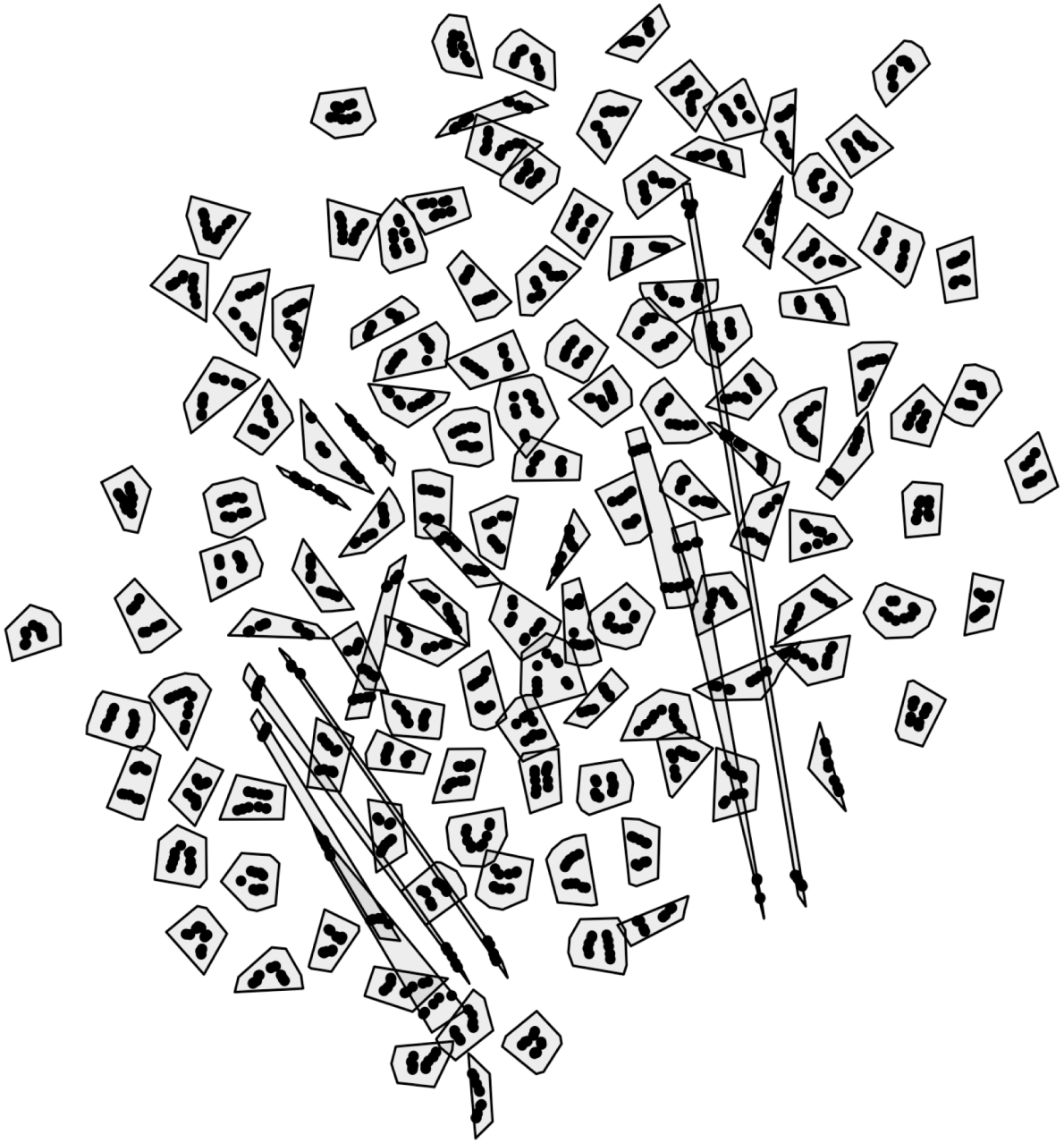
Data to be analyzed

- stimulus images input into DCNN
- analyses done on top-layer DCNN output



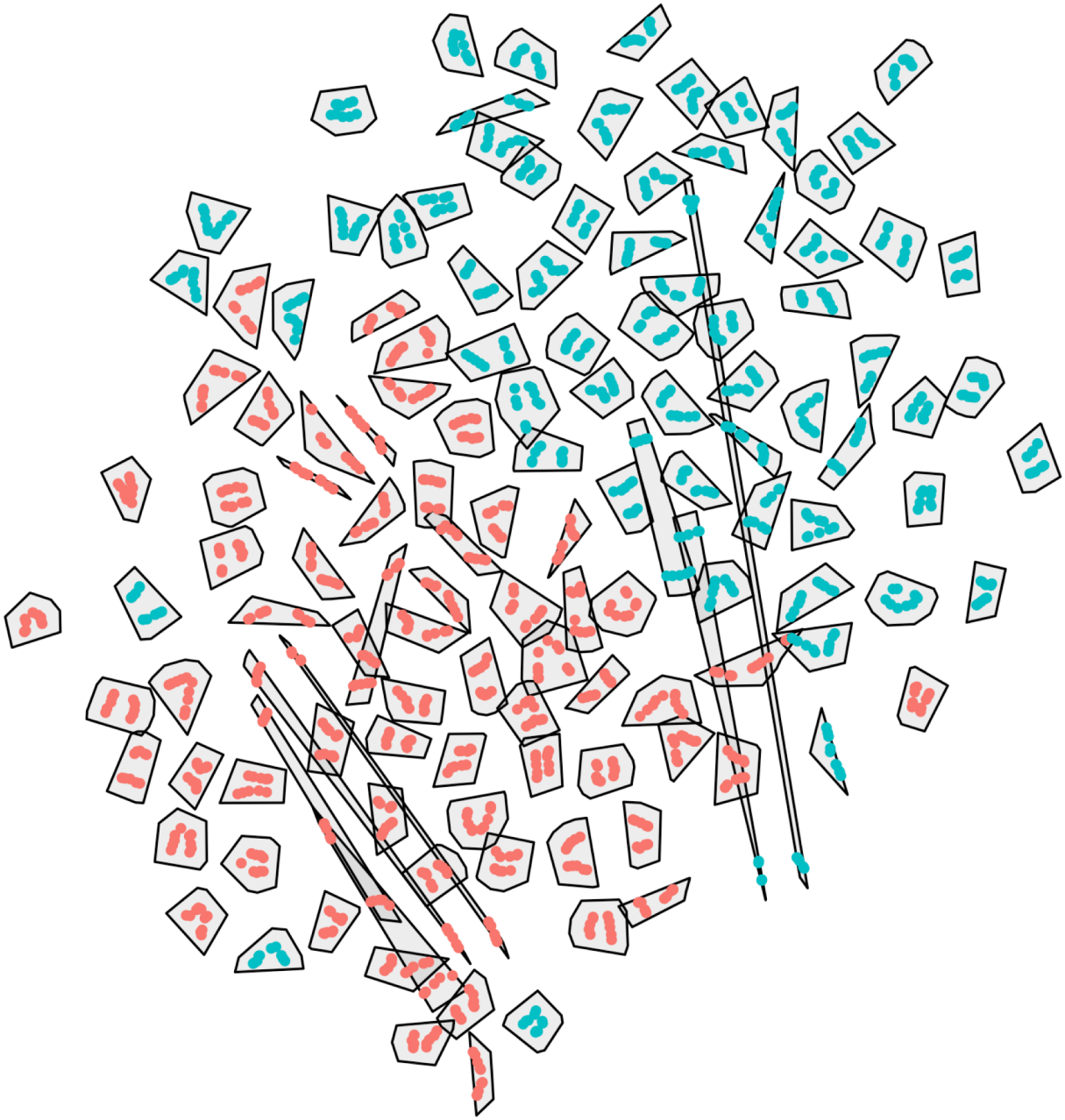
Face space visualizations

- same data coded to see:
  - identity
  - gender
  - illumination
  - viewpoint

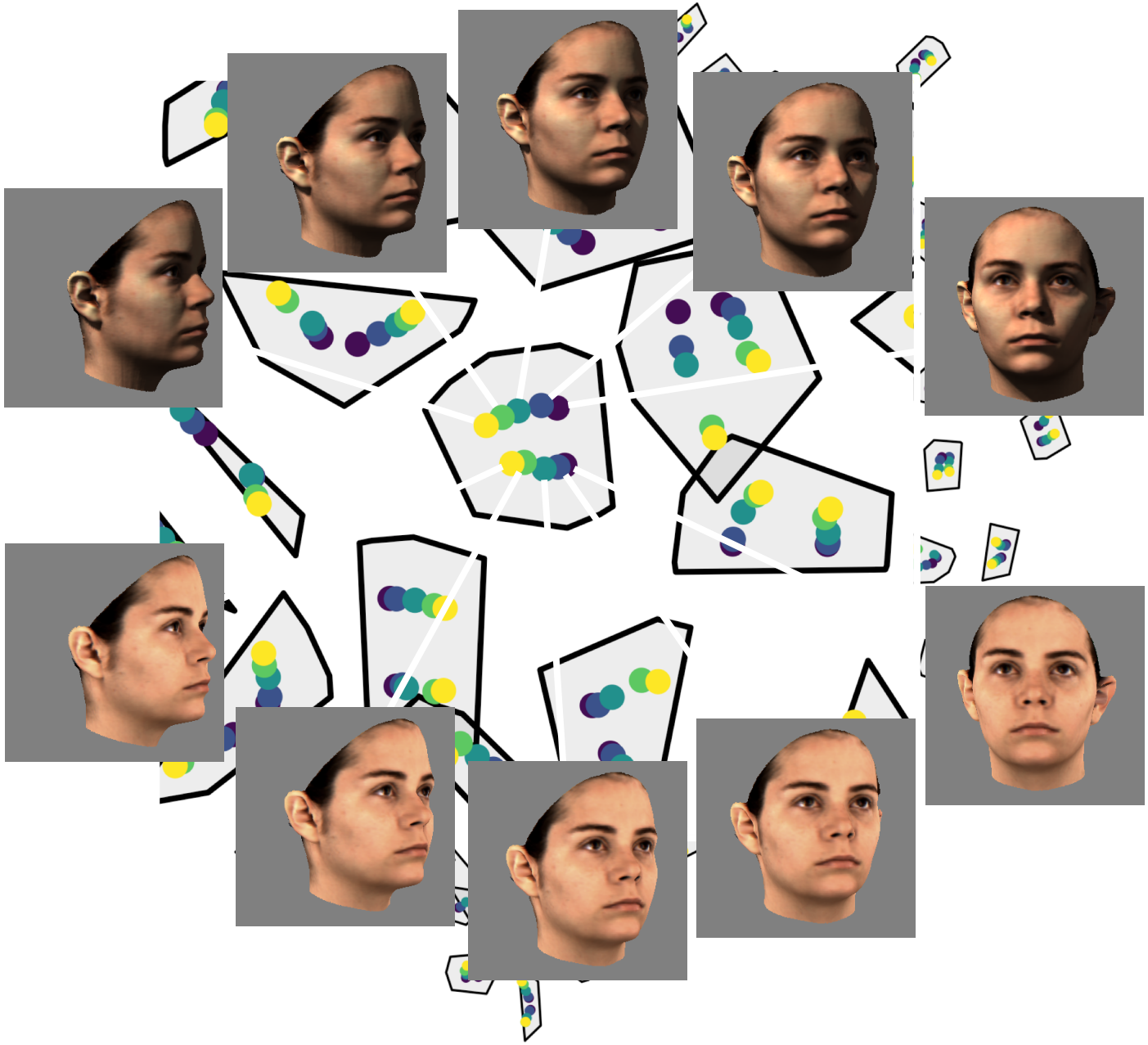


IDENTITY

face identification accuracy:
AUC = 0.997


GENDER
female
male

ILLUMINATION
- ambient
- spotlight



VIEW
- 0° (frontal)
- 20°
- 30°
- 45°
- 60°

Meta-data Prediction

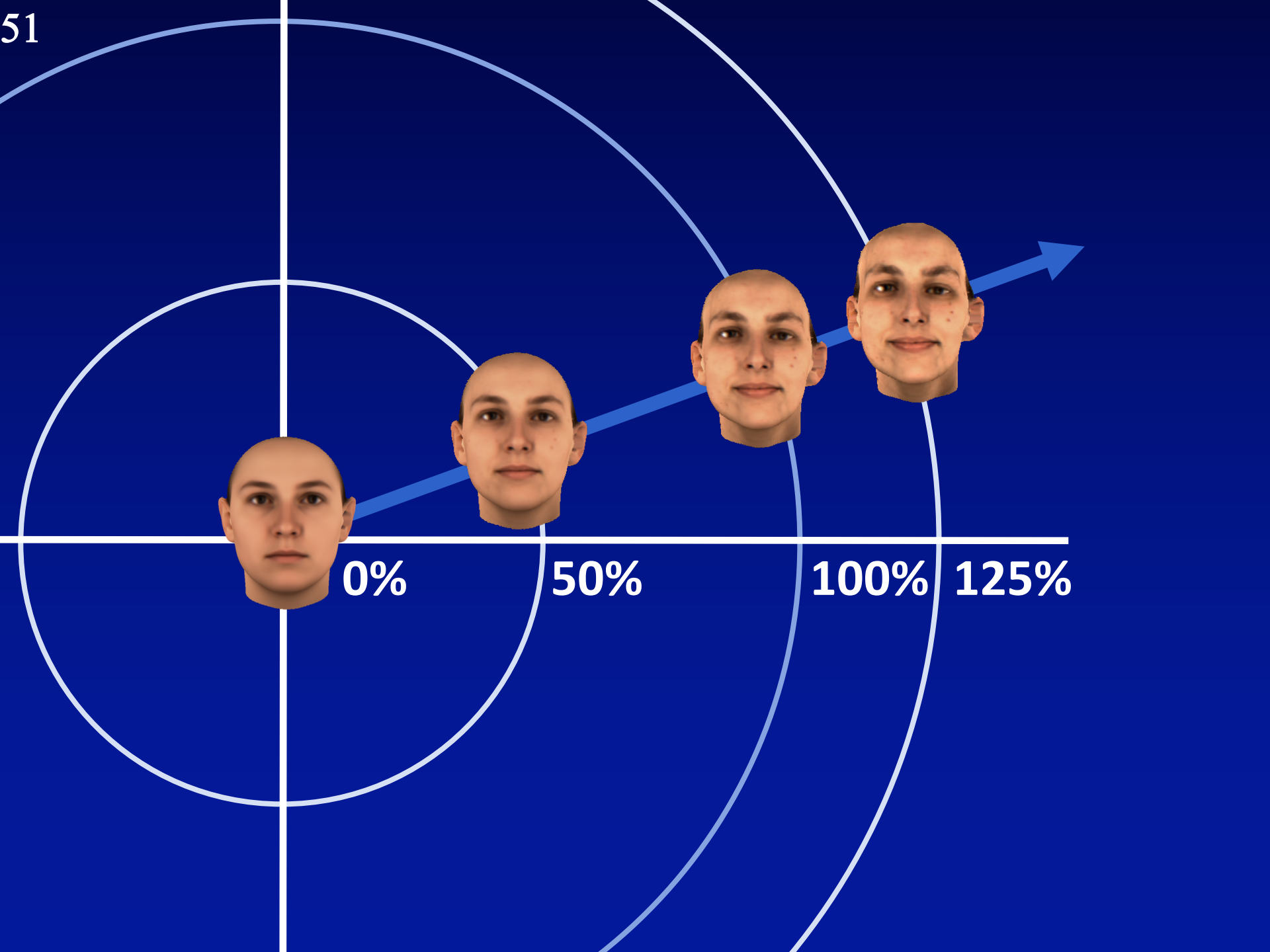
- gender
  - percent correct: 90.98%, $p < .01$
- illumination
  - percent correct: 97.44%, $p < .01$
- viewpoint prediction
  - average error: 7.97° ($SD = 6.02$) , $p < .01$

So far…

- DCNN trained for identity discrimination retains information about non-identity variables
  - gender
  - viewpoint
  - illumination
- hierarchically organized
- image and identity information
  - can be read-out linearly linearly

Understanding Facial Identity

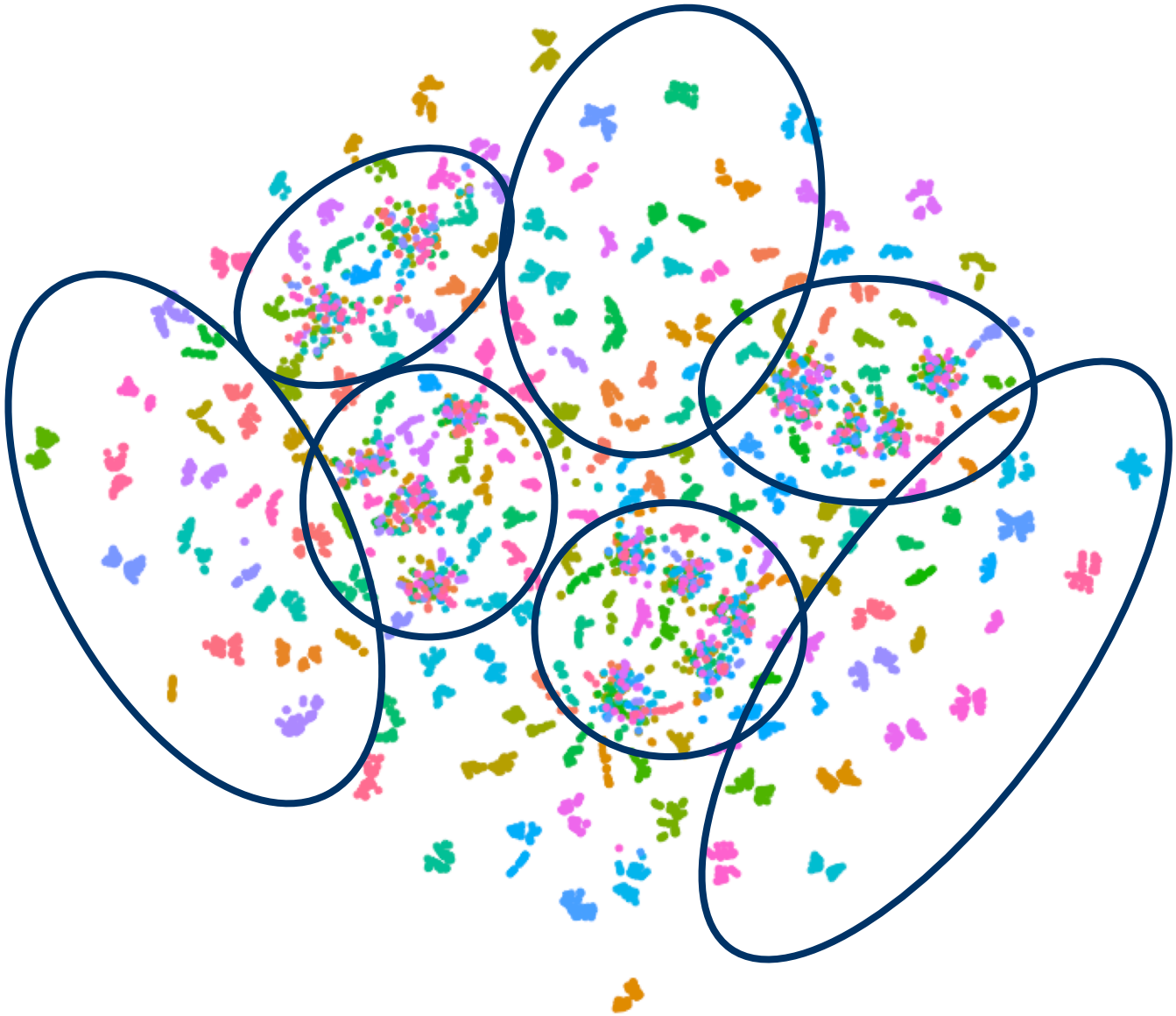
- Manipulate facial distinctiveness
  - 3D Morphable model or Active Appearance Model
  - Morph laser scans (Blanz & Vetter, 1999)


0%
50%
100%
125%

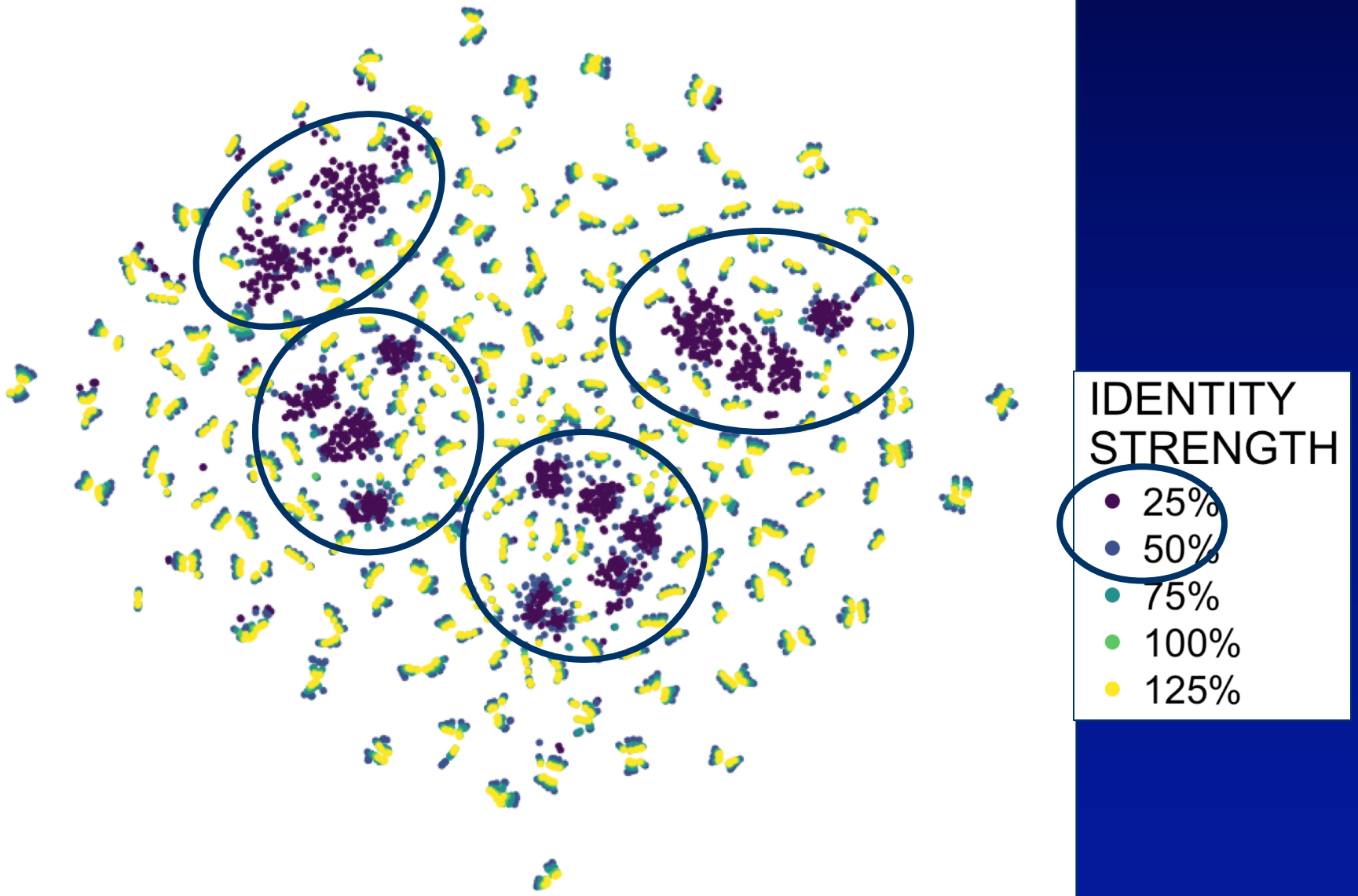
Identity Strength Variation

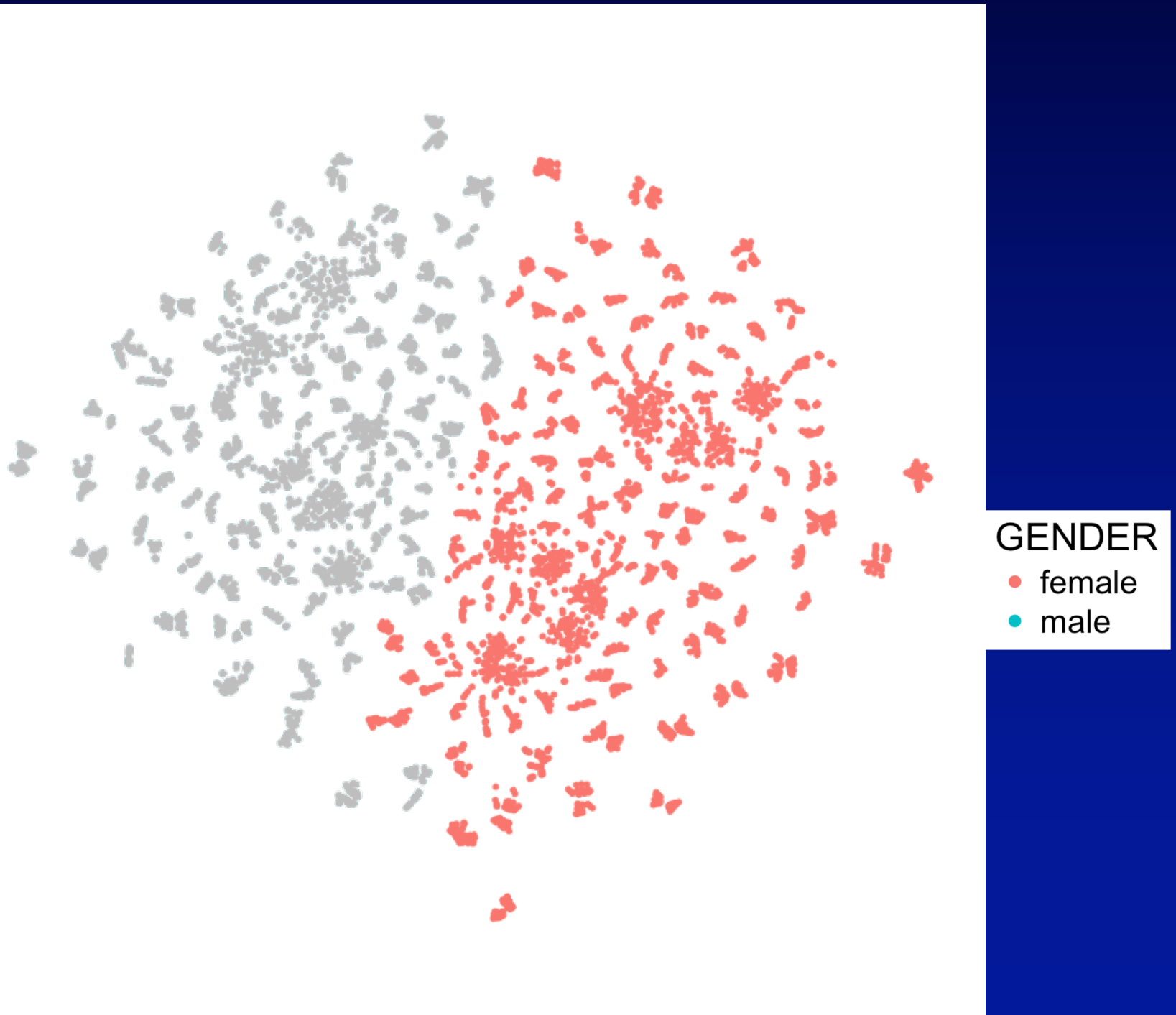
25% **50%** **75%** **100%** **125%**

(veridical) **(caricature)**

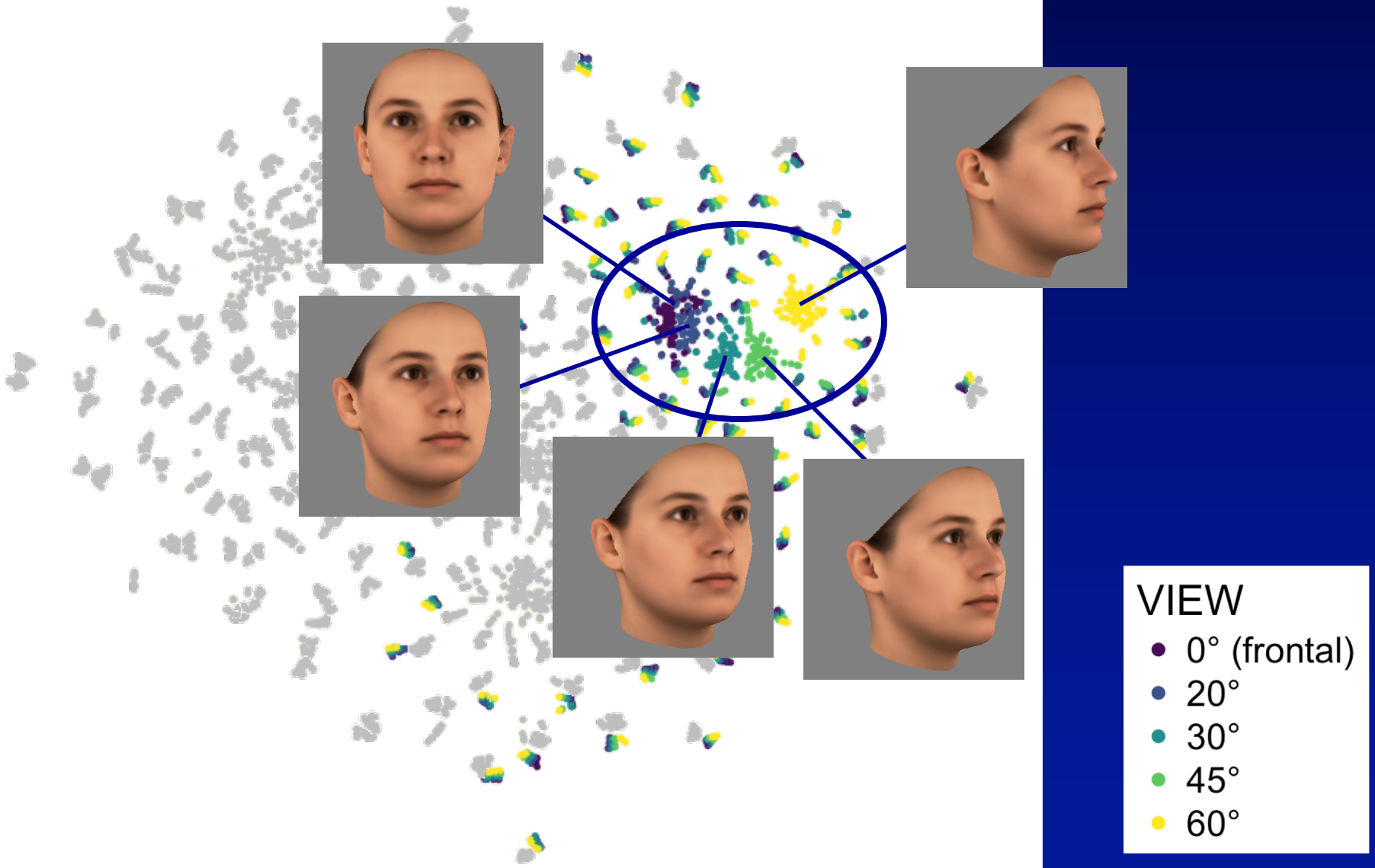


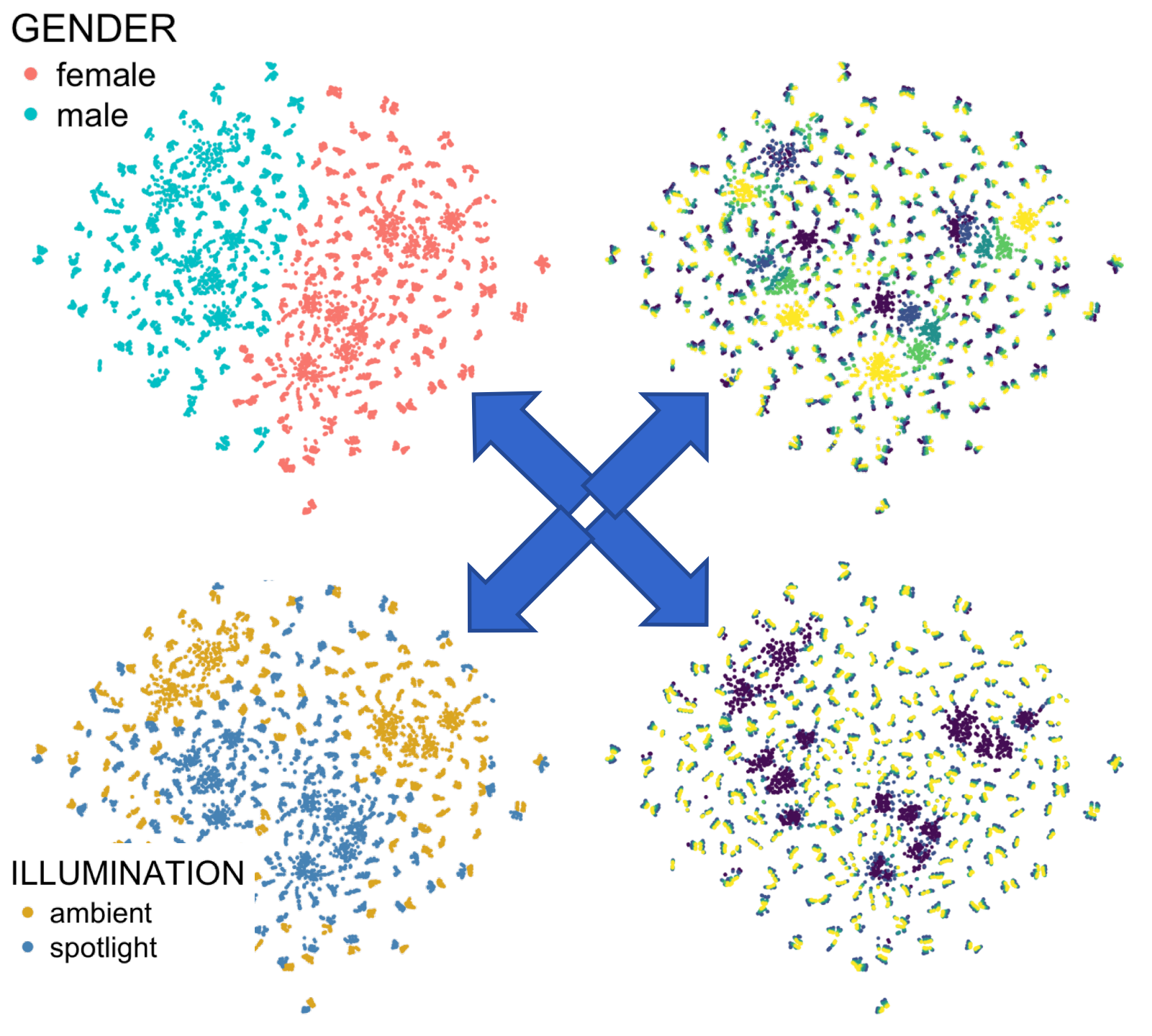
Identity


IDENTITY
STRENGTH
25%
50%
75%
100%
125%


GENDER
female
male


ILLUMINATION
ambient
spotlight


VIEW
0° (frontal)
20°
30°
45°
60°


GENDER
female
male
VIEW
0° (frontal)
20°
30°
45°
60°
IDENTITY STRENGTH
25%
50%
75%
100%
125%
ILLUMINATION
ambient
spotlight

Implications

- DCNN face space -> hierarchically nested pattern of similarity
- *gender* > *identity* > *illumination* > *viewpoint*
- arises spontaneously from network trained for identity
- shows coding flexibility when identity info is weak!
- Other variables??? Demographics etc.?
  - *Future work!*

Thank You!

Acknowledgements. Work supported in part by the Intelligence Advanced Research Projects Activity (IARPA). This research is based upon work supported by the Office of the Director of National Intelligence (ODNI), Intelligence Advanced Research Projects Activity (IARPA), via IARPA R&D Contract No. 2014-140716000012. The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements, either expressed or implied, of the ODNI, IARPA, or the U.S. Government. The U.S. Government is authorized to reproduce and distribute reprints for Governmental purposes notwithstanding any copyright annotation thereon.

O’Toole Lab

12/12/18

Matt Hill

Connor Parde

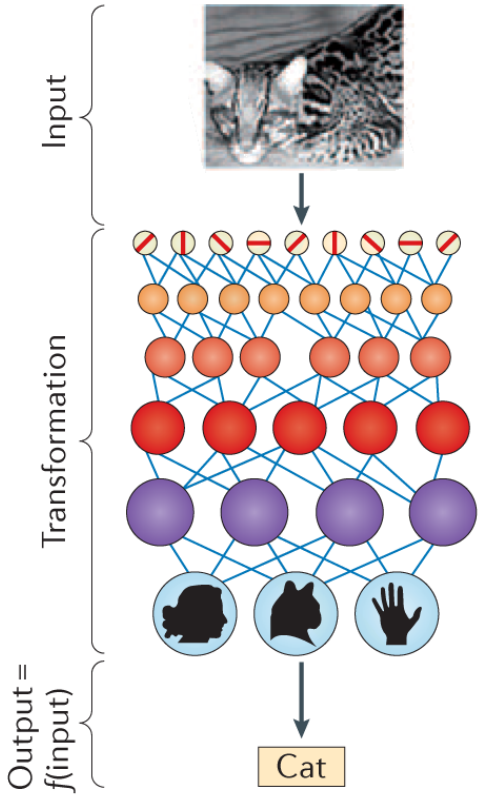
Ivette Colon

Thank you!

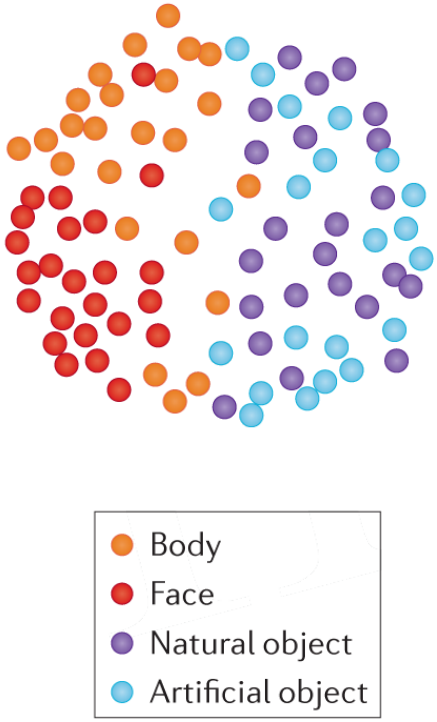
Carlos Castillo

To understand the role of ventral temporal cortex (VTC) in categorization we adapted Marr's framework to modern neuroscience

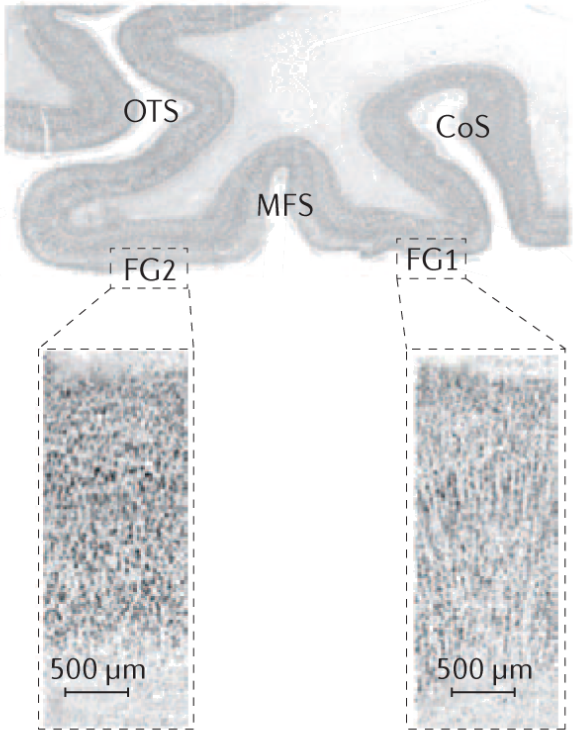
Computations: What are the computational goals of VTC?



Representations: How does VTC represent information to support computations?



Implementation: How are representations physically implemented in VTC?



Grill-Spector & Weiner, Nature Reviews Neuroscience 2014

What are these features?

- Not generic across categories (mostly)
  - *athletic, masculine*
- single features encompass multiple spatial scales and multiple shapes
- constellations of features
- *map easily onto words that entail constellations of features in a non-visual way?*

Deep convolutional neural networks: Face space

Parde, Hill, Colon, Castillo, Chen, Sankaranayraran & O'Toole, A. J. (2017). *Proceedings of the IEEE Conference on Automatic Face and Gesture Recognition*

O'Toole, Castillo, Parde, Hill & Chellappa (2018). Face space representations in deep convolutional neural networks. *Trends in Cognitive Science*.

Hill, Parde, Castillo, Ranjan, Chen, Blanz & O'Toole, (*almost submitted*). Nested hierarchical representations of faces and images in deep convolutional neural networks

O’Toole Lab

12/12/18

Acknowledgements

- This research is based upon work supported by the Office of the Director of National Intelligence (ODNI), Intelligence Advanced Research Projects Activity (IARPA), via IARPA R&D Contract No. 2014-14071600012. The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements, either expressed or implied, of the ODNI, IARPA, or the U.S. Government. The U.S. Government is authorized to reproduce and distribute reprints for Governmental purposes notwithstanding any copyright annotation thereon.

Approach

- visualize the space of face representations
  - highly structured
- analyze information retained in the feature codes
  - classifiers to predict viewpoint, illumination, etc.
- probe the network with controlled images and visualize representations of image variables
  - Hill et al. (almost submitted!).